

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

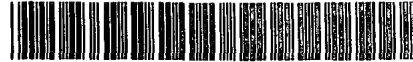
**IMAGES ARE BEST AVAILABLE COPY.**

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

(19)



Europäisches Patentamt  
European Patent Office  
Office européen des brevets



(11) Publication number:

**0 637 152 A1**

(12)

**EUROPEAN PATENT APPLICATION**(21) Application number: **93480104.4**(51) Int. Cl.<sup>8</sup>: **H04L 12/56**(22) Date of filing: **30.07.93**(43) Date of publication of application:  
**01.02.95 Bulletin 95/05**(94) Designated Contracting States:  
**DE FR GB**(71) Applicant: **INTERNATIONAL BUSINESS  
MACHINES CORPORATION**  
Old Orchard Road  
Armonk, N.Y. 10504 (US)(72) Inventor: **Alfonsi, Jean-Pierre**  
Chemin des Combes  
F-06610 La Gaude (FR)  
Inventor: **Lebizay, Gérald****113, avenue des Pollus****F-06140 Vence (FR)**Inventor: **Galand, Claude****56, avenue des Tuillères****F-06800 Cagnes Sur Mer (FR)**Inventor: **Maurel, Olivier****15, rue Forville****F-06110 Le Cannet (FR)**(74) Representative: **de Pena, Alain**  
**Compagnie IBM France**  
**Département de Propriété Intellectuelle**  
**F-06610 La Gaude (FR)**(54) **Method and apparatus to speed up the path selection in a packet switching network.**

(57) Currently, the routing algorithms compute all the available paths in the network, from the source node to the destination node before to select the optimal route. No assumption is made on the network topology and the route computation is often time and resource consuming. Some paths which are calculated are not acceptable due to the particular geographical configuration of the network.

In the real world, large transport networks are not fully meshed. The present invention is based on the observation that networks are usually built around a hierarchical structure. A set of nodes, interconnected by high throughput lines, are used to build a 'Backbone' (401) with a high degree of meshing to allow the redundancy and reliability required by the user. The other nodes or 'local' nodes (404) are attached to one or several backbone nodes. It is the network designer responsibility, at the configuration time to define for each node what is its attribution : backbone (402) or local node (404). The list of the node attributions appears in the topology table (306) and is updated each time a node is added to or dropped from the network. The routing algorithm can take advantage of the particular network topology to drastically reduce the complexity of paths computation. For a given connection, only a limited number of nodes are eligible and are taken in account by the algorithm in the optimal route search. The object of the invention is to split the network in backbone and local nodes to speed up the path selection.

**EP 0 637 152 A1**

# NETWORK DECOMPOSITION

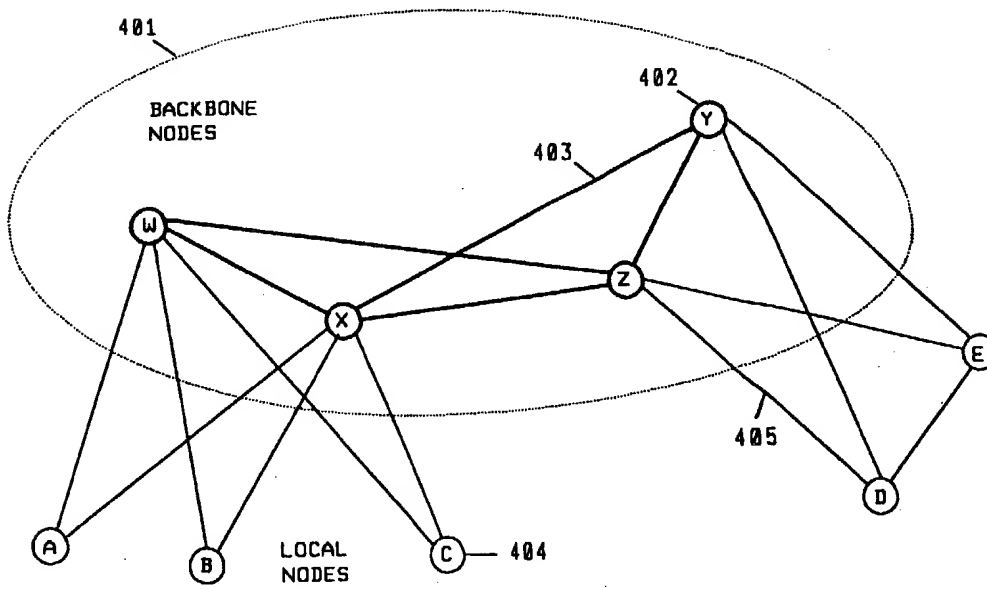


FIGURE 4

**Technical field**

The present invention relates to a high speed transmission system in a large packet switching network, more particularly, to an apparatus and method for speeding up the path selection between a source and a destination node and specially a method to decompose the network into backbone and local nodes.

**Background art**

It has become increasingly useful to interconnect a plurality of data processing elements by means of a packet switching network in which data is transmitted as data assemblages called "packets". Such networks include a plurality of interconnected switching nodes which, in turn, are connected to end nodes supporting data processing equipments. Such packet networks can become quite large with an extensive geographical distribution. In such a situation, the selection of an efficient path between two end nodes which wish to communicate with each other becomes of paramount importance. Different methods are summarized by H. Nussbaumer in *Teleinformatique II* (pages 92 to 117) Presses Polytechniques Romandes 1987.

**Path Selection**

The general problem which must be solved at the network level is to find a path between a source and a destination node. When the network is using datagrams, the path selection must be done individually for each packet. For virtual circuits, the path decision is done once only at the time of the circuit establishment. In both cases, the choice of a routing algorithm is not easy because it must satisfy a large number of often conflicting requirements. This algorithm must be simple to implement to not complicate the nodes realization, it must ensure a correct progress of the packets whatever the disturbance on the network. The algorithm must be able to provide satisfying results despite of the traffic variations and the network topology. It must also ensure the equity between the different end users and provide an equal distribution of the rights to access the network. At last, the routing algorithm must allow, when possible, to exploit the network in an optimal way, according to a criterion which can vary with the utilization type. In most of the cases, the network is realized in a way to minimize the packet transit time and to transfer the maximum number of packets. The principal objectives are respectively the reduction of the transit time and the increase of the throughput. In other cases, the objective is to decrease the communication cost, or to develop a reliable network able to operate correctly (without too sensible performance degradation) either in case of catastrophic line, node failure or peaks of traffic. Because of the variety of the constraints, there are a large number of different routing types. Some correspond to deterministic or adaptive policies according to their capacity of integrating the traffic variations and the network topology. Routing algorithms can be centralized if the paths are determined from a particular node. Others are distributed between all nodes this is favorable in a reliability point of view but the algorithm is more complicated and the path optimization is more difficult to implement. Some Algorithms are difficult to classify: they are using special techniques called Flooding Routing or Random Routing.

**Flooding Routing**

The Flooding Routing is based on a very simple principle consisting for each node to retransmit packets on all output links of the node, except, but the input link. A node connected to K other nodes retransmit K-1 copies of the packet which has been just received. This method allows the delivery to the destination node of at least one copy of the packet with one condition : it must exist at least one path between the source to the destination node. This routing is ensured even when the network topology is changing for example after a catastrophic failure of some components in the network. The Flooding Routing allows a very good robustness in the network exploitation. Furthermore, as all possibles links between the source and the destination node are tried in an exhaustive way, the method ensures that at least one copy of the packet will reach the recipient through the shorter path with the minimum delay if the network is not overloaded. This technique is simple to implement since neither routing tables to route the messages, nor knowledge of the geographical position of the destination node in the network are required. The recipient must only know that the packet is addressed to it.

The price to pay for these qualities of robustness, simplicity and rapidity is a bad utilization of the network resources and a tendency to the congestion. Usually, it is advantageous to place in the packet header a field indicating the maximum node number that a packet is authorized to pass. This field is decremented each time the packet go through a node, and packets with an empty field are discarded.

### Random or Stochastic Routing

This technique, like the Flooding Routing, does not demand to the nodes to have the complete knowledge of the network structure and of the traffic to take a routing decision at their own level. However, here, to limit the generation of a too important ghost traffic, the nodes avoid to send systematically over all output links copies of the packets they receive. The Random Routing consists in emitting one or more counterparts of the received packet on output links selected at random. This method is also called Selective Flooding. This policy leads to a very simple routing at the node level and limits the saturation of the network due to the packet duplication. In return, the average routing delay is longer than this resulting from deterministic techniques. Packets are taking sinuous routes instead of taking the most direct path toward the destination node.

### Deterministic Routing

#### Fixed Routing

The Fixed Routing defines paths that must take the different packets according to the general network characteristics such as the network topology and the mean traffic expected on the communication lines. The routing rules are established once and aims for optimizing the performance criterion privileged by the user. In most of the cases, the average packet transmission time through the network must be minimized. The method consists in building a routing table at the level of each node. Its consultation allows the node to determine on which output link the packet it has just received must be transferred. The Fixed Routing is very simple in its principles. The only processing done in the nodes consists in queues management and tables reading and the optimization algorithm is initiated only once at the network design level. The Fixed Routing aims for a long term and a global optimization of the network, but in comparison with the random routing, this technique allows to considerably speed up the packet transmission through the network.

### Adaptive Routing

Contrary to the Fixed Routing, the purpose of the Adaptive Routing is to satisfy at any time the optimization criterion. The tables are permanently updated according to, for example, the instantaneous state of the traffic on the lines.

- Centralized Routing

When the characteristics of the network fluctuate, it is possible to adapt the routing by assigning to one node the responsibility to update periodically the routing tables according to the traffic variations and the topology. This method, simple in its principles, is called Centralized Routing. Its principal disadvantage is to generate an important auxiliary traffic and to subordinate the good functioning of the network to only one node. The different nodes must send to the routing center the state of the network as they perceive it (operational adjacent nodes, traffic ...), and the routing center must, in its turn, dispatch to the nodes their routing tables. The auxiliary traffic is concentrated around the routing center and this has for result to saturate this part of the network. Furthermore, the Centralized Routing can generate some problems at the time the tables are refreshed because said tables cannot be received at the same time by all the nodes. A solution is to decentralize the tables establishment at the level of each node.

- Local Routing

The local and distributed routing methods are both based on the establishment by each node of its own routing table according to information locally collected. With the local routing technique, each node builds its table without exchanging information with the neighboring nodes. In its most simple form, the method consists in placing the packet just received in the shorter output queue and in transmitting it as rapidly as possible. The local algorithm tends in its principle to circulate the packets very quickly in the network. However, as the routes are selected in some way or other the mean paths length is far to be minimal.

- Distributed Routing

The Distributed Routing is a local method in which neighboring nodes are exchanging messages concerning the traffic and the network condition to update their routing and delay tables.

- Hierarchical Routing

With the non Stochastic Routing methods, each node requires a table with as many rows as nodes in the network and a number of columns equal to the number of output links. Furthermore, when the

routing is adaptive, nodes must exchange periodically messages to update their routing tables. The tables size and the importance of the control messages are increasing rapidly with the number of nodes and become unacceptable when the network contains more than ten nodes. To solve this problem, the network is divided in a hierarchy of sub networks as illustrated in Figure 0. In particular, optimal clustering structures (003) are determined so as to minimize the length of the routing tables required. Said tables are taking into account only the nodes in a sub network (003) and a packet addressed to a node in another sub network will necessarily go through the access node (005) of this remote sub network. Packets transit through some mandatory nodes (005) and the price to pay for this simplification is that smaller routing tables give less precise routing information which then results in longer path lengths for the message traffic. Hierarchical Routing procedures are particularly effective for large packet switched networks (001). With smaller routing tables, they require less storage and processing in the nodes (002) as well as less communications overhead in the lines (004).

The problem of the optimization of the hierarchical levels has been studied by **L. Kleinrock and F. Kamoun - Hierarchical Routing for Large Networks, Computer Networks, Vol. 1, pp. 155-174, January 1977.**

The main idea for reducing the routing table length is to keep, at any node, complete routing information about nodes which are close to it and less information about nodes located further away from it. This can be realized by providing one entry per destination for the closer nodes and one entry per set of destinations for the remote nodes.

The reduction of routing table length is achieved through a hierarchical partitioning of the network. Basically, an m-level hierarchical clustering of a set of nodes (Figure 0) consists in grouping the nodes (002) into a 1st level clusters (003), which in turn, are grouped into 2nd level cluster etc... This operation continues in a bottom up fashion. The mth level cluster is the highest level cluster (3rd level cluster) and as such it includes all the nodes of the network (001).

Unfortunately, the gains in table length are accompanied with an increase of the message path length in the network. This result is a degradation of the network performance (delay, throughput) due to the excess internal traffic caused by longer path length. However, in very large networks, an enormous table reduction may be achieved with no significant increase in network path length.

### High Performance Networks

The data transmission is now evolving with a specific focus on applications and by integrating a fundamental shift in the customer traffic profile. Driven by the growth of workstations, the local area networks (LAN) interconnection, the distributed processing between workstations and super computers, the new applications and the integration of various and often conflicting structures - hierarchical versus peer to peer, wide (WAN) versus local (LAN) area networks, voice versus data - the data profile has become higher in bandwidth, bursting, non deterministic and requires more connectivity. Based on the above, it is clear that there is strong requirement for supporting distributed computing applications across high speed networks that can carry LAN communications, voice, video, and traffic among channel attached hosts, business, engineering workstations, terminals, and small to intermediate file servers. This vision of a high speed multiprotocol network is the driver for the emergence of fast packet switching networks architectures in which data, voice, and video information is digitally encoded, chopped into small packets and transmitted through a common set of nodes and links.

### Throughput

The key requirement of these new architectures is to reduce the end-to-end delay in order to satisfy real time delivery constraints and to achieve the necessary high nodal throughput for the transport of voice and video. Increases in link speeds have not been matched by proportionate increases in the processing speeds of communication nodes and the fundamental challenge for high speed networks is to minimize the packet processing time within each node.

As example, for meeting a typical 100 ms delay to deliver a voice packet between two end users :

- A total of 36 ms might be needed for the packetization and play-out functions at the end points.
- About 20 ms is the unalterable propagation delay needed, say, to cross the United States.
- There remains 44 ms for all the intra-node processing time as the packet moves through the network. In a 5 nodes network, each node would have about 8 ms for all processing time including any queueing time. In a 10 nodes network, each node would have about 4 ms.

Another way of looking the same constraint is illustrated in Figure 1: taking a node with an effective processing rate of 1 MIPS (Millions of Instructions Per Second), it is possible to fill a 9.6 kbps line with 1000 byte packets even if a network node must execute 833 000 instructions per packet processed. For a 64 kbps line the node can afford 125 000 instructions per packet. In order to fill an OC24 link, however, our 1 MIPS node could only execute 7 instructions per packet. In the latter case even an effective rate of 10-30 MIPS would allow only 70-200 instructions per packet.

In order to minimize the processing time and to take full advantage of the high speed/low error rate technologies, most of the transport and control functions provided by the new high bandwidth network architectures are performed on an end-to-end basis. The flow control and particularly the path selection are managed by the access points of the network which reduces both the awareness and the function of the intermediate nodes.

### Routing Modes

The routing within the network presents two aspects :

1. Determining what the route for a given connection shall be,
2. Actually switching the packet within a switching node.

There are many methods of determining a route through a network. For very high throughput, the critical item is that the switching elements must be able to decide where to route an incoming packet in a very short portion of time. As described in the document *High Speed Networking Technology, An Introductory Survey (pages 88 to 96) - GG24-3816-00 ITSC Raleigh March 1992*, the routing modes, the most widely used, are the Source Routing and the Label Swapping.

- Source Routing

The Source Routing is a particular implementation of the Distributed Routing for connectionless networks. The source node (or access node) is responsible for calculating the route the packet must take through the network. A routing field is appended to every packet sent and that field is used by the intermediate nodes to direct the packet towards its destination. In this technique, the sending node must know the network topology and it must be able to find the optimal route. But once the route is determined, intermediate nodes do not need to refer to any tables or parameters to make the routing decision. The next stage of the route is right there in the packet header. A drawback of this method is that the routing field in the packet header takes some storage and is an overhead. But this is quite small and the benefits of being able to make a fast routing decision out-weigh the small increase in bandwidth overhead. Since all routing information is contained in each packet a set up is not required along the path to provide an end to end communication. Thus, data applications which benefit from a datagram service can be effectively supported by this technique. However, the datagram traffic is characterized by the fact that each datagram can be viewed as a connection and requires the computation of a path. Each packet is routed through the network as a self contained unit and is independent of all other packets. The key point for the source node is to determine for each datagram the optimal route in the shorter lapse of time.

- Label Swapping

The Label Swapping is a particular implementation of the Distributed Routing for connection oriented networks. These networks typically multiplex many connections (or sessions) on a link using some form of logical "channel". Each packet sent on the link has a header which includes an arbitrary number identifying which logical connection that this packet belongs to. For example, each packet can contain in its routing field a label that will be used by the intermediate nodes to identify the transmission link the packet should be routed over. When a packet arrives at a node, the label is extracted from its header, and used to index a routing table that provides a new label along with a link address. The new label is over written on the old label, and the packet is forwarded using the link address. The information in the routing table is updated at the connection set up time. Each node on a path selects a label for each direction of the connection, and sends it to the neighboring node.

The call set up and the resource reservation process comprises the following steps:

- **A connection request** is specified by the user via a set of parameters including origin and destination network address and data flow characteristics,
- **The path determination** is realized by the source node from its local routing topology database
- **The route reservation** is sent out in a special message along the specified nodes. Said message, which is usually routed according to the previously described Source Routing technique, signals the nodes to set up their connection tables and to reserve their resources to provide the level of service required by the traffic type (for example, a bandwidth reservation on each of the links of the

path).

- **The tables update** reflects the changes in the availability of the network resources. The information is sent to every node in the network.

The Label Swapping technique requires that the connection tables be set up and maintained dynamically. That means that when a new connection is established or an old one is terminated the tables are updated (the database of network topology can of course be maintained quite separately). This is possible if the connection set up is relatively infrequent and is not too time critical. This restriction makes datagram transport quite inefficient. However, connections that require very short packets, like real-time voice connections, can be effectively supported by this technique, because of the low packet overhead. Once the connection is established, there is no need to place a destination address in the header every time a packet is sent. All is needed is a field to specify which connection is to be used for this packet.

### Path Selection Criterion

One of the major criterion for selecting paths between nodes in packet networks is the minimum hop count and minimum path length. The hop count is the number of links used to construct the path between the two end nodes. The path length is a function of the overall transmission delay imposed by the path between the two end nodes. In most high speed networks, the delay (path length) is not a major consideration since the worst-case delay through such networks is nearly always acceptable. The hop count, however, is a direct measure of the amount of resources required to implement a given path and hence is of considerable importance in selecting paths. It is to be noted that a selected path need not be a minimum hop count path since congestion on the network links may force the choice of a larger hop count path. However, such longer alternate paths cannot be allowed to grow without limit since inordinate amounts of network resources might be committed to one path, resulting in further congestion for other paths and forcing yet longer hop count paths to be selected for yet other connections. The long term network throughput could thereby be adversely affected. The problem, then, is to select a path between an origin node and a destination node which has a minimum hop count, a minimum path length, which does not utilize an inordinate amount of network resources.

### Key requirements

The Distributed Routing mechanism in large and high speed packed switching networks supporting both connection oriented and connectionless routing modes implies some requirements in terms of performance and resource consumption which can be summarized as follows :

- the source node (or the node providing the route calculation for the source node) must be able to decide where to route an incoming packet in a very short portion of time (the computation must be sufficiently rapid to compute an optimum path for each connection request)
- the switching time in the intermediate nodes must be minimized, (minimum processing time)
- the network resources along the selected path must be optimized according to the criterion of the minimum hop count.
- Control messages must be as much as possible limited not to overload the network.

In high speed networks the cost for updating the routing tables generated by the exchange of control messages is not critical so long as this can be performed before the connection set up. The very high line throughput authorizes the interchanging of routing information without creating a significant burden on the communication lines themselves. These considerations suggest a better access and use of the routing tables maintained within the nodes without sacrificing the criterion of optimal path contrary to the Hierarchical Routing method proposed by L. Kleinrock and F. Kamoun.

### Summary of the Invention

Currently, the routing algorithms compute all the available paths in the network, from the source node to the destination node before to select the optimal route. No assumption is made on the network topology and the route computation is often time and resource consuming. For example, some paths which are calculated are not acceptable due to the particular geographical configuration of the network.

In the real world, large transport networks are not fully meshed. The present invention is based on the simple observation that networks are usually built around a hierarchical structure. A set of nodes, interconnected by high throughput lines, are used to build a 'Backbone' with a high degree of meshing to



allow the redundancy and reliability required by the user. The other nodes or 'local' nodes are attached to one or several backbone nodes. It is the network designer responsibility, at the configuration time to define for each node what is its attribution : backbone or local node. The list of the node attributions appears in the topology table and is updated each time a node is added to or dropped from the network. The routing algorithm can take advantage of the particular network topology to drastically reduce the complexity of paths computation. For a given connection, only a limited number of nodes are eligible and are taken in account by the algorithm in the optimal route search.

The object of the invention is to split the network in backbone and local nodes to speed up the path selection.

#### **Description of the drawings**

Figure 0 represent a network topology for Hierarchical Routing according to the method suggested in the prior art by L. Kleinrock and F. Kamoun

Figure 1 shows the processing times (or number of instructions per second) required in function of the different line throughputs supported by the high speed networks.

Figure 2 shows a typical model of high speed packet switching network including the access and transit nodes claimed in the present invention.

Figure 3 describes a high speed Routing Point according to the present invention.

Figure 4 illustrates the decomposition of the network in a plurality of backbone nodes and local nodes according to the present invention.

Figure 5 illustrates the addition in the network of new nodes and links and their classification in the topology database according to the present invention.

Figure 6 represents the optimum path tree of node A

Figure 7 represents the optimum path tree of node A according to the present invention.

Figure 8 summarizes the path selection methodology according to the present invention.

#### **Description of the preferred embodiment of the invention**

As illustrated in Figure 2, a typical model of communication system is made of several user networks (212) communicating through a high performance network (200) using private lines, carrier provided services, or public data networks. Each user network can be described as a set of communication processors and links (211) interconnecting large computers used as enterprise servers (213), user groups using workstations or personnel computers attached on LAN (Local Area Networks 214), applications servers (215), PBX (Private Branch eXchange 216) or video servers (217). These user networks, dispersed in different establishments, need to be interconnected through wide area transport facilities and different approaches can be used for organizing the data transfer. Some architectures involve the checking for data integrity at each network node, thus slowing down the transmission. Others are essentially looking for a high speed data transfer and to that end the transmission, routing and switching techniques within the nodes are optimized to process the flowing packets towards their final destination at the highest possible rate. The present invention belongs essentially to the latter category and more particularly to the fast packet switching network architecture detailed in the following paragraphs.

#### **High speed packet switching networks**

The general view in Figure 2 shows a fast packet switching transmission system comprising eight nodes (201 to 208) each node being interconnected by means of high speed communication lines called Trunks (209). The access (210) to the high speed network by the users is realized through Access Nodes (202 to 205) located at the periphery. These Access Nodes comprise one or more Ports, each one providing an access point for attaching external devices supporting standard interfaces to the network and performing the conversions required to transport the users data flow across the network from and to other external devices. As example, the Access Node 202 interfaces respectively a Private Branch eXchange (PBX), an application server and a hub through three Ports and communicates through the network by means of the adjacent Transit Nodes 201, 208 and 205.

### Switching Nodes

Each network node (201 to 208) includes a Routing Point where the incoming data packets are selectively routed on the outgoing Trunks towards the neighboring Transit Nodes. Such routing decisions are made according to the information contained in the header of the data packets. In addition to the basic packet routing function, the network nodes also provide ancillary services such as :

- the determination of routing paths for packets originated in the node,
- directory services like retrieving and updating information about network users and resources,
- the maintaining of a consistent view of the physical network topology, including link utilization information, and
- the reservation of resources at access points of the network.

Each Port is connected to a plurality of user processing equipments, each user equipment comprising either a source of digital data to be transmitted to another user system, or a data sink for consuming digital data received from another user system, or, typically, both. The interpretation of the users protocols, the translation of the users data into packets formatted appropriately for their transmission on the packet network (200) and the generation of a header to route these packets are executed by an Access Agent running in the Port. This header is made of Control and Routing Fields.

- The Routing Fields contain all the information necessary to route the packet through the network (200) to the destination node to which it is addressed.
- The Control Fields include, among other things, an encoded identification of the protocol to be used in interpreting the Routing Field (Source Routing, Label Swapping, ....).

### Routing Points

Figure 3 shows a general block diagram of a typical Routing Point (300) such as it can be found in the network nodes (201 to 208) illustrated in Figure 2. A Routing Point comprises a high speed packet Switch (302) onto which packets arriving at the Routing Point are entered. Such packets are received :

- from other nodes over high speed transmission links (303) via Trunk Adapters (304).
- from users via application adapters called Ports (301).

Using information in the packet header, the adapters (304, 301) determine which packets are to be routed by means of the Switch (302) towards a local user network (307) or towards a transmission link (303) leaving the node. The adapters (301 and 304) include queuing circuits for queuing packets prior to or subsequent to their launch on the Switch (302).

The Route Controller (305) calculates the optimum routes through the network (200) so as to satisfy a given set of quality of service specified by the user, and to minimize the amount of network resources used to complete a communication path. Then, it builds the header of the packets generated in the Routing Point. The optimization criterion includes the number of intermediate nodes, the characteristics of the connection request, the capabilities and the utilization of the Trunks in the path ... All the information necessary for the routing, about the nodes and transmission links connected to the nodes, are contained in a Network Topology Database (306). Under steady state conditions, every Routing Point has the same view of the network. The network topology information is updated when new links are activated, new nodes added to the network, when links or nodes are dropped or when link loads change significantly. Such information is originated at the network node to which the resources are attached and is exchanged by means of control messages with all other Route Controllers to provide the up-to-date topological information needed for route calculation (such database updates are carried on packets very similar to the data packets exchanged between end users of the network). The fact that the network topology is kept current in every node through continuous updates allows dynamic network reconfigurations without disrupting end users logical connections (sessions).

The incoming transmission links to the packet Routing Point may comprise links from external devices in the local user networks (210) or links (Trunks) from adjacent network nodes (209). In any case, the Routing Point operates in the same manner to receive each data packet and forward it on to another Routing Point as dictated by the information in the packet header. The fast packet switching network operates to enable a communication between any two end user applications without dedicating any transmission or node facilities to that communication path except for the duration of a single packet. In this way, the utilization of the communication facilities of the packet network is optimized to carry significantly more traffic than would be possible with dedicated transmission links for each communication path.

**Network Decomposition****Modified Bellmann-Ford Algorithm**

There are several possible algorithms to compute an optimal route in a network. The modified Bellmann-Ford algorithm is one of the most currently used in fast packet packet switching networks. As described in European Patent Application 93480030.1 - *Methods and Apparatus for Optimum Path Selection in Packet Transmission Networks*, this one is invoked for each connection with the following parameters :

- the source node address,
- the destination node address,
- the quality of service required for the connection. For example :
  - maximum call set up delay (very important parameter for connections serviced in real time),
  - maximum end to end transit delay,
  - maximum information loss, error probability,
  - ...

The algorithm uses the local copy, in the source node (or access node), of the network topology to determine the minimum hop and path length with the destination node which meets the quality of service requirements. As mentioned, the modified Bellman-Ford algorithm makes no assumption on the network geographical configuration and it requires the same complexity whether the network is fully meshed or hierarchized. The purpose of the present invention is, for a given connection, to simplify the network topology by reducing the number of eligible nodes for the path calculation.

**Algorithm Complexity**

The total hop count necessary from a given node to reach all the other nodes gives a measure of the complexity  $C$  of the path selection according to the minimum hop criterion.  $C$  is proportional to the average number of links  $L$  per node, to the total number of nodes  $N$  in the network and to the average number of optimal hops.

$$\begin{aligned}
 C = N \times & \quad 0 \\
 & 1 \cdot 1 (N-1)/N + \quad (\text{first hop}) \\
 & \quad 1 \\
 & 2 \cdot 1 (N-1)/N + \quad (\text{second hop}) \\
 & \quad 2 \\
 & 3 \cdot 1 (N-1)/N + \quad (\text{third hop}) \\
 & \quad \dots \dots \dots \\
 & \quad i \\
 & (i+1) \cdot 1 (N-1)/N \quad (i+1 \text{ th hop})
 \end{aligned}$$

$$\begin{aligned}
 C &= \text{hopt} \cdot N \cdot 1 \\
 &= \text{hopt} \cdot N \cdot 2L/N = \text{hopt} \cdot 2L
 \end{aligned}$$

with :

- hopt = average hop count for an optimal path
- $L$  = average number of links per node
- $N$  = total number of nodes in the network
- $L$  = total number of links in the network

The complexity is directly related to the total number of links in the network. For a given path search between two nodes, the complexity can be reduced in a very large proportion by pre-selecting the links that will be used for the computation of the route. This pre-selection is efficient if the minimum hop number

criterion is not degraded in the same proportion, that means if **hopt** remains constant.

### Nodes and Links Attributes

5 The basic idea of the present invention is to pre-select in a given node the physical links which must be considered for building a path to another node and then to use said pre-selected links at path selection time. This invention details also, how to initiate and maintain the link table.

Some nodes are selected to built a backbone. The other nodes called local nodes are attached to at least one node of this backbone. Both local and backbone nodes are able to support Ports and Trunks, 10 without any restriction. The node attributes are recorded in the topology database and updated for each change in the network, node addition or node suppression. Based on the node attributes, each link is qualified by a new attribute, according to the following rules:

Node_i	Node_j	Link_ij Type
Backbone	Backbone	Backbone
Backbone	Local	Local
Local	Backbone	Local
Local	Local	Local

20 These link attributes are not defined by the network designer. They are dynamically built by each node, according to the Node\_i and Node\_j types.

The classification of the nodes can be done according to one of the following methods:

- by hand in case of small networks. The search of the path from the source to the destination node is limited to the backbone level, all possible search through the local network are eliminated.
- by using off-line a path selection algorithm for each possible source-destination.

### Topology Database

30 The topology database is replicated in every node and contains information on the network nodes and links. A topology algorithm is used to maintain a consistent view of the network in all nodes. The topology database contains two classes of information :

- The physical topology of the network including static information like physical characteristics of nodes and links,
- The link utilization.

Only the first class of parameters are related to the present invention.

### Assumptions

- The network comprises N nodes and L links,
- Every node has the same knowledge of the network topology.

### Definitions

45 In a given node (Node\_i), it must be built a simple structure able to determine which links have to be used to reach any other node (Node\_j). This matrix structure can be defined as follows :

- There is a line for each link l, and a column for each node of the network. The so defined structure is a L x N matrix.
- The matrix element, Eln, is a Boolean value :
  - Eln = 1 (TRUE) means : Link\_l can be used to reach Node\_n from Node\_i
  - Eln = 0 (FALSE) means : Link\_l cannot be used to reach Node\_n from Node\_i

### Matrix Initiation

55 The matrix is defined, at the configuration time, independently in each node, according to the following principles :

- Links (or matrix lines)

- The links going from Node\_\_i to the backbone nodes are all usable to reach other nodes (except Node\_\_i itself),
  - The links between any backbone nodes are all usable to reach any destination in the network.
  - The other local links (not directly attached to Node\_\_i) are usable only to reach the local node(s) they are attached to.
- 5
- Nodes (or matrix columns). In Node\_\_j column,
    - All backbone links are usable,
    - All links attached to Node\_\_i or/and Node\_\_j are usable
    - All other links are not usable.
- 10

### Matrix Update

Each time a new node or link is added, the topology database located in every node of the network must be updated. This is done by means of control messages containing the new topology and particularly the attributes of the new node. The matrix is updated according to the following rules :

15

- The addition of a new local Node\_\_k involves the addition of a new column with a :
    - value 1 (TRUE) for backbone links and links attached to Node\_\_i and/or Node\_\_k.
    - value 0 (FALSE) for all other local links
  - The addition of a new backbone Node\_\_k involves the addition of a column with a :
    - value 1 (TRUE) for the backbone links, and links attached to Node\_\_i.
    - value 0 (FALSE) for all other local links
  - The addition of a new local link Ljk involves the addition of a new row with a :
    - VALUE 1 (TRUE) for the local node j or k if that link is not directly attached to Node\_\_i,
    - value 1 (TRUE) for all nodes (except Node\_\_i) if that link is directly attached to Node\_\_i,
    - value 0 (FALSE) for all other nodes.
  - the addition of a new backbone link Ljk involves the addition of a new row in the matrix with a 1 value (TRUE) for all nodes (except Node\_\_i).
- 20
- 25

### Optimal Path Search Methodology

- 30
- The method according to the present invention can be summarized as follows (Figure 8) :
- **Step 1** : The node is characterized with an attribute : the node is defined as backbone or local node.
  - **Step 2** : The node stores the network configuration in its topology database. This database is initiated and maintained by means of control messages exchanges between all the nodes in the network.
  - 35 • **Step 3** : From the information stored in the topology database the node identifies the backbone and the local nodes.
  - **Step 4** : The node determines the attribute of each link according to the node attributes.
  - **Step 5** : The node selects the usable links for each destination node in the network by building a L x N matrix.
  - 40 • **Step 6** : At each connection request, the routing algorithm is computed among the set of pre-selected usable links.
  - **Step 7** : During all the time of the connection, the data packets are sent to the destination node along the path previously computed.

### Example

As illustrated in Figure 4, the network is decomposed at the configuration time, in a plurality of backbone (401) and local nodes :

Backbone nodes (402) : W,X,Y,Z  
 50 Local nodes (404) : A,B,C,D,E  
 and  
 Backbone links (403) : Lwx,Lwz,Lxy,Lxz,Lyz,  
 Local Links (405) : Law,Lax,Lbw,Lbx,Lcw,Lcx, Ldy,Ldz,Lde,Ley,Lez  
 The corresponding matrix for the node A is :

55

	Nodes	A	B	C	D	E	W	X	Y	Z
	Links									
5	Law	0	1	1	1	1	1	1	1	1
	Lax	0	1	1	1	1	1	1	1	1
	Lbw	0	1	0	0	0	0	0	0	0
	Lbx	0	1	0	0	0	0	0	0	0
10	Lcw	0	0	1	0	0	0	0	0	0
	Lcx	0	0	1	0	0	0	0	0	0
	Ldy	0	0	0	1	0	0	0	0	0
15	Ldz	0	0	0	1	0	0	0	0	0
	Lde	0	0	0	0	0	0	0	0	0
	Ley	0	0	0	0	1	0	0	0	0
20	Lez	0	0	0	0	1	0	0	0	0
	Lwx	0	1	1	1	1	1	1	1	1
	Lwz	0	1	1	1	1	1	1	1	1
	Lxy	0	1	1	1	1	1	1	1	1
25	Lxz	0	1	1	1	1	1	1	1	1
	Lyz	0	1	1	1	1	1	1	1	1

30 Figure 5 shows the example of the addition in the network of two nodes : a local node F and a backbone node T with the :

new backbone links : Ltx ,Lty,Ltw

and new local links: Lfx,Lfz,Lat,Lab

The matrix is updated accordingly as illustrated hereunder :

35

40

45

50

55

# EP 0 637 152 A1

Nodes	A	B	C	D	E	W	X	Y	Z	F	T
Links											
5 Law	0	1	1	1	1	1	1	1	1	1	1
Lax	0	1	1	1	1	1	1	1	1	-1	1
Lbw	0	1	0	0	0	0	0	0	0	0	0
Lbx	0	1	0	0	0	0	0	0	0	0	0
10 Lcw	0	0	1	0	0	0	0	0	0	0	0
Lcx	0	0	1	0	0	0	0	0	0	0	0
Ldy	0	0	0	1	0	0	0	0	0	0	0
15 Ldz	0	0	0	1	0	0	0	0	0	0	0
Lde	0	0	0	0	0	0	0	0	0	0	0
Ley	0	0	0	0	1	0	0	0	0	0	0
20 Lez	0	0	0	0	1	0	0	0	0	0	0
Lwx	0	1	1	1	1	1	1	1	1	1	1
Lwz	0	1	1	1	1	1	1	1	1	1	1
Lxy	0	1	1	1	1	1	1	1	1	1	1
25 Lxz	0	1	1	1	1	1	1	1	1	1	1
Lyz	0	1	1	1	1	1	1	1	1	1	1
30 Lfx	0	0	0	0	0	0	0	0	0	1	0
Lfz	0	0	0	0	0	0	0	0	0	1	0
Ltx	0	1	1	1	1	1	1	1	1	1	1
Lty	0	1	1	1	1	1	1	1	1	1	1
35 Ltw	0	1	1	1	1	1	1	1	1	1	1
Lat	0	1	1	1	1	1	1	1	1	1	1
40 Lab	0	1	1	1	1	1	1	1	1	1	1

To simplify the notation, the matrix in Node\_\_i is represented as a L entry table  $T_i(k)$  ( $k = 1, \dots, L$ ) with :

- $T_i(k) = 1$  when link k is usable for any path,
- $T_i(k) = 0$  when link k is not usable for any path,
- $T_i(k) = j$  when the link k is usable only on the path from Node\_\_i to Node\_\_j ( $j = 1, \dots, N$  except i)).

In our example, the tables  $T_i$  in nodes A to Z have the following values :

	Links	TA	TB	TC	TD	TE	TF	TT	TW	TX	TY	TZ
5	Law	1	A	A	A	A	A	A	A	A	A	A
	Lax	1	A	A	A	A	A	A	A	A	A	A
	Lbw	B	1	B	B	B	B	B	B	B	B	B
	Lbx	B	1	B	B	B	B	B	B	B	B	B
10	Lcw	C	C	1	C	C	C	C	C	C	C	C
	Lcx	C	C	1	C	C	C	C	C	C	C	C
	Ldy	D	D	D	1	D	D	D	D	D	D	D
15	Ldz	D	D	D	1	D	D	D	D	D	D	D
	Lde	0	0	0	E	D	0	0	0	0	0	0
	Ley	E	E	E	E	1	E	E	E	E	E	E
20	Lez	E	E	E	E	1	E	E	E	E	E	E
	Lwx	1	1	1	1	1	1	1	1	1	1	1
	Lwz	1	1	1	1	1	1	1	1	1	1	1
	Lxy	1	1	1	1	1	1	1	1	1	1	1
25	Lxz	1	1	1	1	1	1	1	1	1	1	1
	Lyz	1	1	1	1	1	1	1	1	1	1	1
30	Lfx	F	F	F	F	F	1	F	F	F	F	F
	Lfz	F	F	F	F	F	1	F	F	F	F	F
	Ltx	1	1	1	1	1	1	1	1	1	1	1
	Lty	1	1	1	1	1	1	1	1	1	1	1
35	Ltw	1	1	1	1	1	1	1	1	1	1	1
	Lat	1	A	A	A	A	A	A	A	A	A	A
40	Lab	B	A	0	0	0	0	0	0	0	0	0

The path between two nodes in the network is considered as optimum if the number of hops is minimum. All optimum paths create a tree of which root is the source node. Figure 6 shows all the optimum paths established from node A. At the first level are placed all nodes adjacent to the source node, then at the second level all the adjacent nodes to the first level nodes and not already placed and so on until exhaustion. Figure 7 illustrates the path search from node A to node E according to the present invention. As previously described, the eligible links to reach node E are defined in the table TA.



Eligible		Non-Eligible	
Links	TA	Links	TA
Law	1	Lbw	B
Lax	1	Lbx	B
Lat	1	Lcw	C
Ley	E	Lcx	C
Lez	E	Ldy	D
Lwx	1	Ldz	D
Lwz	1	Lde	0
Lxy	1	Lfx	F
Lxz	1	Lfz	F
Lyz	1	Lab	B
Ltx	1		
Lty	1		
Ltw	1		

Only the backbone links, and the local links attaching the source node and the destination node to the backbone (701) are taken into account in the search process. The other links (702) are not taking part to the path selection which reduces the complexity of the routing algorithm computation (for example the Bellman-Ford Algorithm). In our example, four paths - AWZE, ATYE, AXZE, AXYE - are satisfying the minimal hop constraint with a number of three hops.

The complexity of the network represented in Figure 5 can be approximated as follows :

$$C = \text{hopt} \cdot 2L = 1.69 \times (2 \times 23) = 77.7$$

with :

- . hopt = 1.69 hops
- . L = 23 links
- . N = 11 nodes
- . I = 4.18 (46/11) links per node

The average hop number is determined from all the optimum (minimum) paths between the nodes :

- **1 hop** : AW, AT, AX, AB, BW, BX, CW, CX, FX, FZ, DE, DZ, DY, EZ, EY, WT, WX, WZ, XZ, XY, XT, TY, ZY.
- **2 hops** : AC, AZ, AY, AF, BC, BZ, BY, BF, BT, CZ, CY, CF, CT, FD, FE, FY, FT, FW, DX, DT, DW, EX, ET, EW, ZT, YW.
- **3 hops** : AE, AD, BD, BE, CD, CE

$$\begin{aligned} \text{hopt} &= (1 \times 23 + 2 \times 26 + 3 \times 6) / (23 + 26 + 6) = 93/55 \\ &= 1.69 \text{ hops per optimum path} \end{aligned}$$

With the assumption that the pre-selection of the usable links according to the present invention does not degrade the minimum hop constraint, the algorithm complexity can, in our example, be nearly reduced in a factor two :

$$C = \text{hopt} \cdot 2L' = 1.69 \times (2 \times 13) = 43.9$$

with :

- . L' = 13 links
- . hopt = 1.69 hops

This reduction of complexity can be much important typical networks.

## Claims

1. A network access node (300) for a packet switching communication network (200) comprising a plurality of nodes (201..208) interconnected with transmission links (209), said node including means for :  
 5     • receiving and transmitting data packets (301, 302, 304),  
      • storing and updating the network configuration (306),  
 and characterized in that it further comprises means for :  
    • storing a pre-selection of usable links to communicate with each destination nodes located in the network,  
 10    • determining, for each connection request, an optimal routing path from said access node to the destination node among said stored pre-selected links.
2. A network access node according to claim 1 comprising, in said storing means (306), means for :  
 15     • being identified as backbone node or local node,  
      • identifying backbone nodes (402) and local nodes (404),  
      • identifying backbone links (403) and local links (405).
3. A network access node according to claim 2 wherein said link pre-selection means includes means for :  
 20     • selecting as usable for the path determination all backbone links,  
      • selecting as usable for the path determination all links attached to the node and/or to the destination node,  
      • not selecting as usable for the path determination all other links.
4. A network access node according to anyone of claims 1 to 3 wherein said routing path determination  
 25 includes means for :  
    • computing a modified Bellmann-Ford algorithm.
5. A method for selecting in an access node (300) a routing path in a packet switching communication network (200) comprising a plurality of nodes (201..208) interconnected with transmission links (209),  
 30 said method involving the steps of :  
    • storing and updating the network configuration,  
 and characterized in that it further comprises the steps of :  
    • storing a pre-selection of usable links to communicate with each destination nodes located in the network,  
 35     • determining, for each connection request, an optimal packet routing path from said access node to the destination node among said stored preselected links.
6. A method for selecting a routing path according to claim 5 wherein said storing step includes the further steps of :  
 40     • being identified as backbone node or local node,  
      • identifying backbone nodes (402) and local nodes (404),  
      • identifying backbone links (403) and local links (405).
7. A method for selecting a routing path according to claim 6 wherein said link pre-selection step includes the further steps of :  
 45     • selecting as usable for the path determination all backbone links,  
      • selecting as usable for the path determination all links attached to the node and/or to the destination node,  
      • not selecting as usable for the path determination all other links.
8. A network access node according to anyone of claims 5 to 7 wherein said routing path determination  
 50 step includes the further step of :  
    • computing a modified Bellmann-Ford algorithm.

# HIERARCHICAL ROUTING

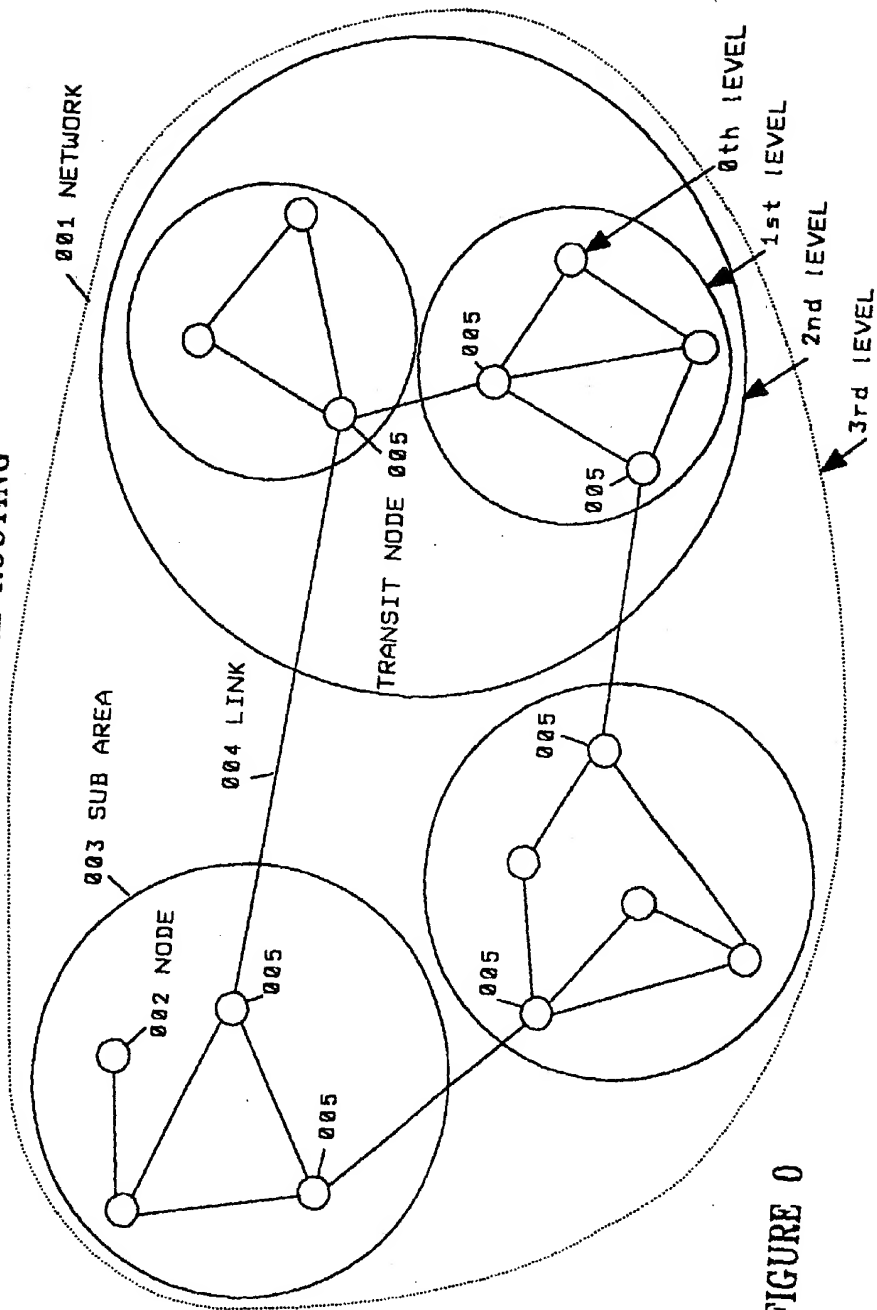


FIGURE 0

# PROCESSING TIME IN TRANSIT NODES

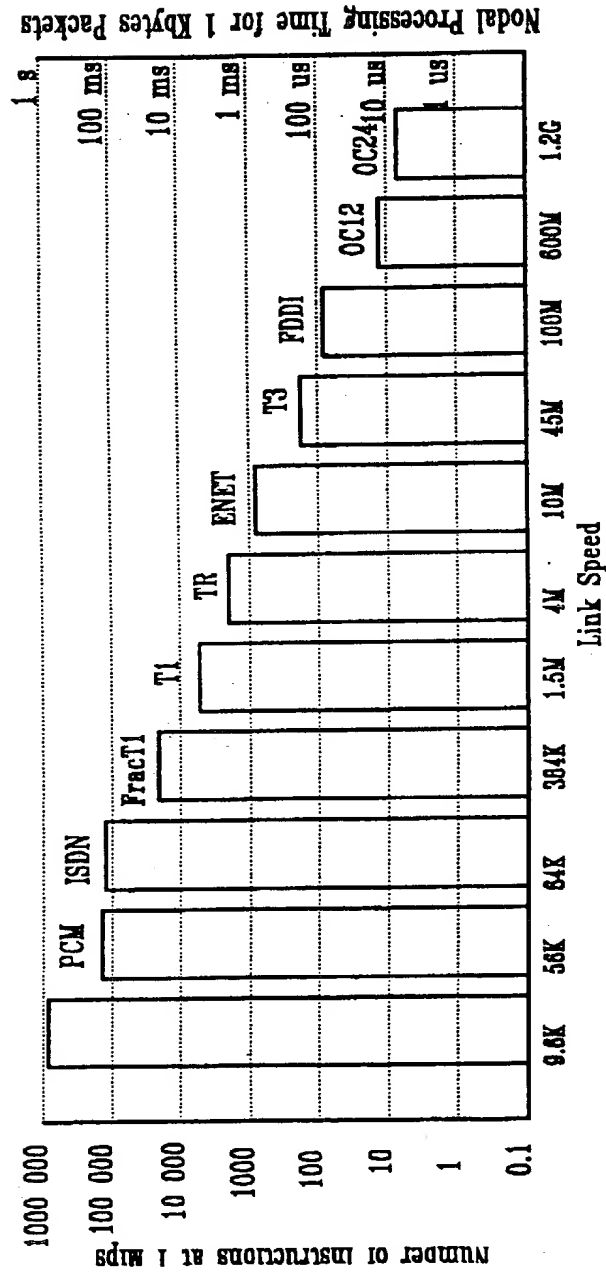


FIGURE 1

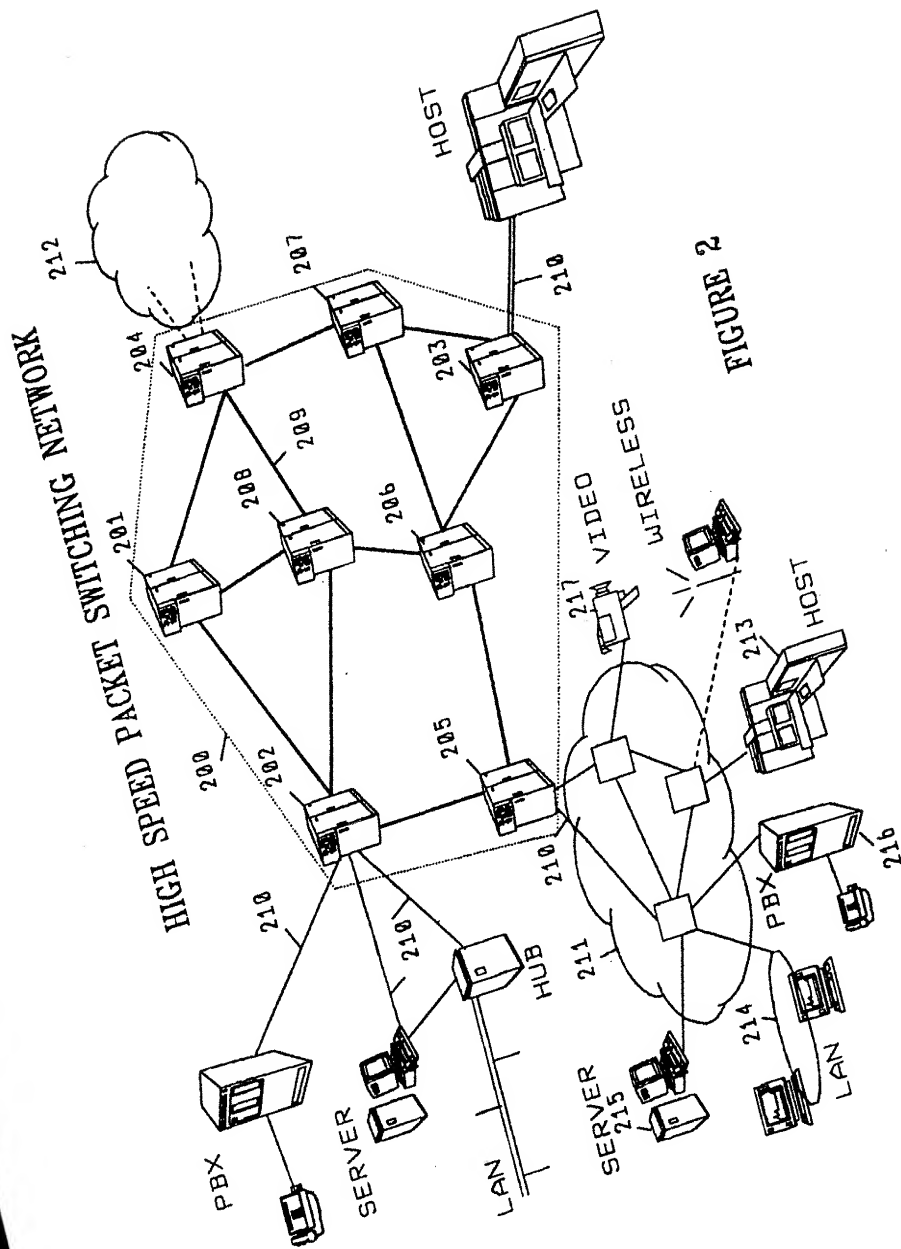


FIGURE 2

# HIGH SPEED PACKET SWITCHING NODE

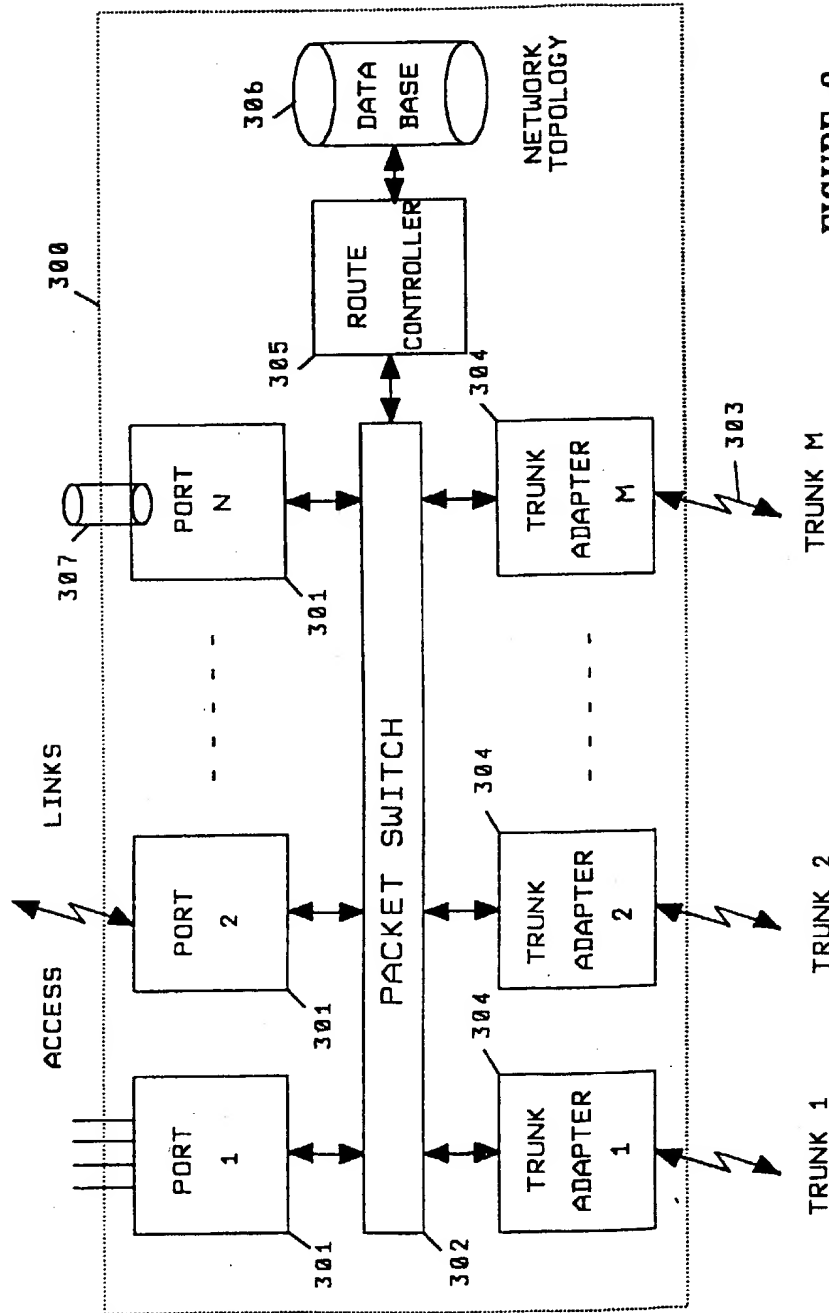


FIGURE 3

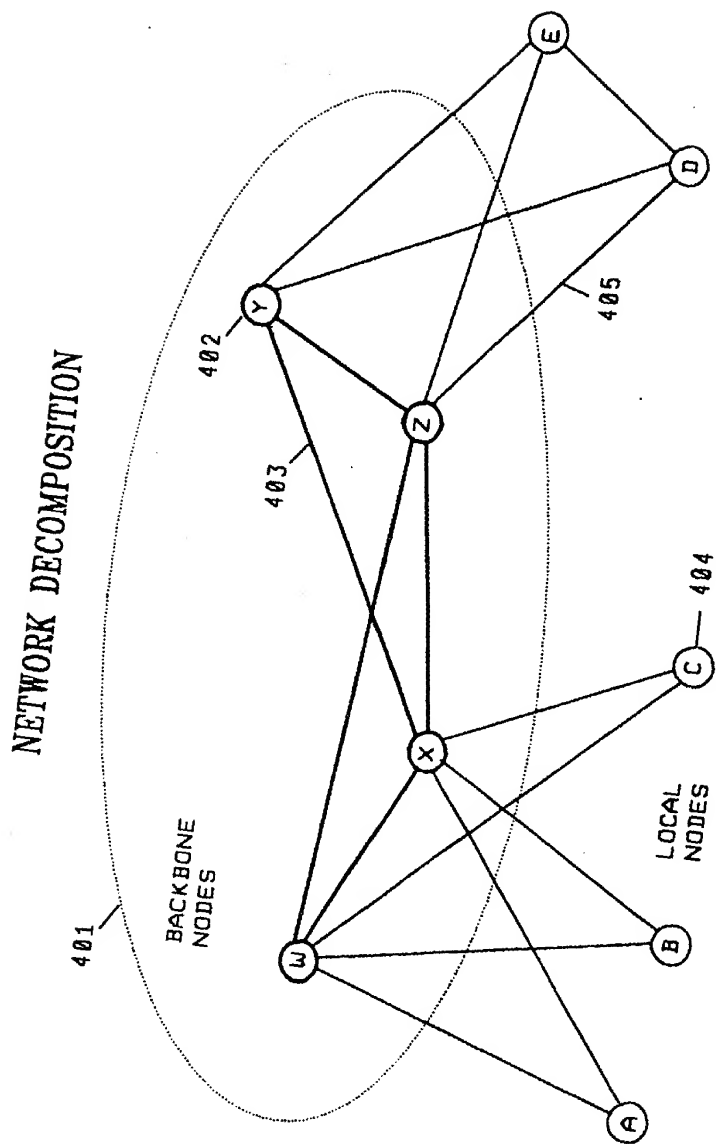


FIGURE 4

# NETWORK UPDATE

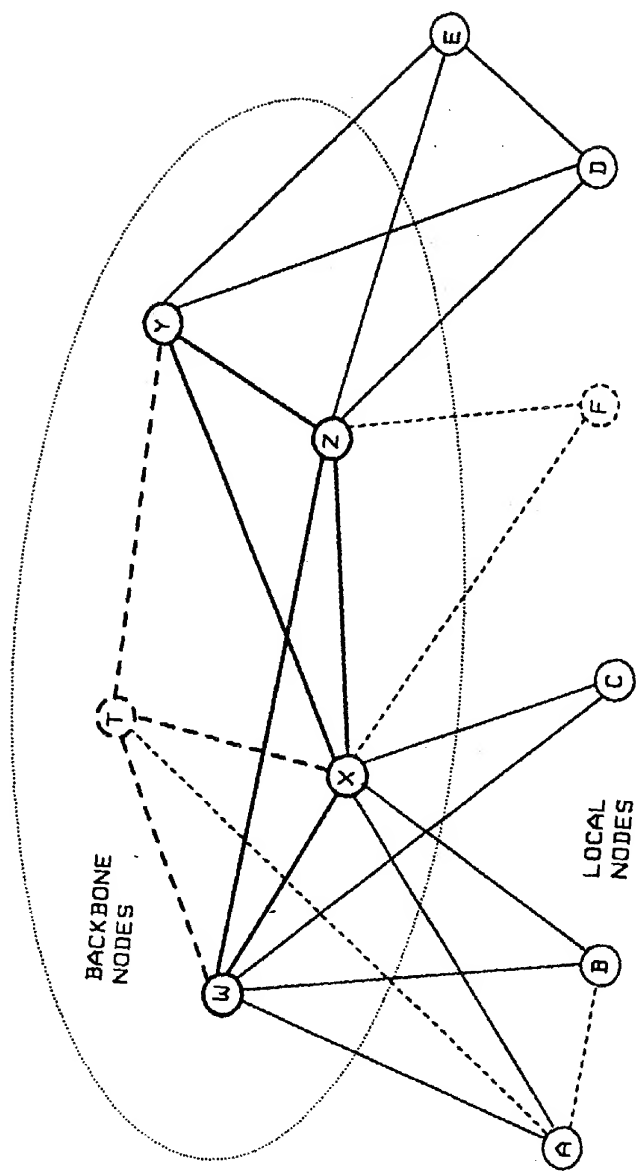


FIGURE 5



Node A - Optimum Paths Tree

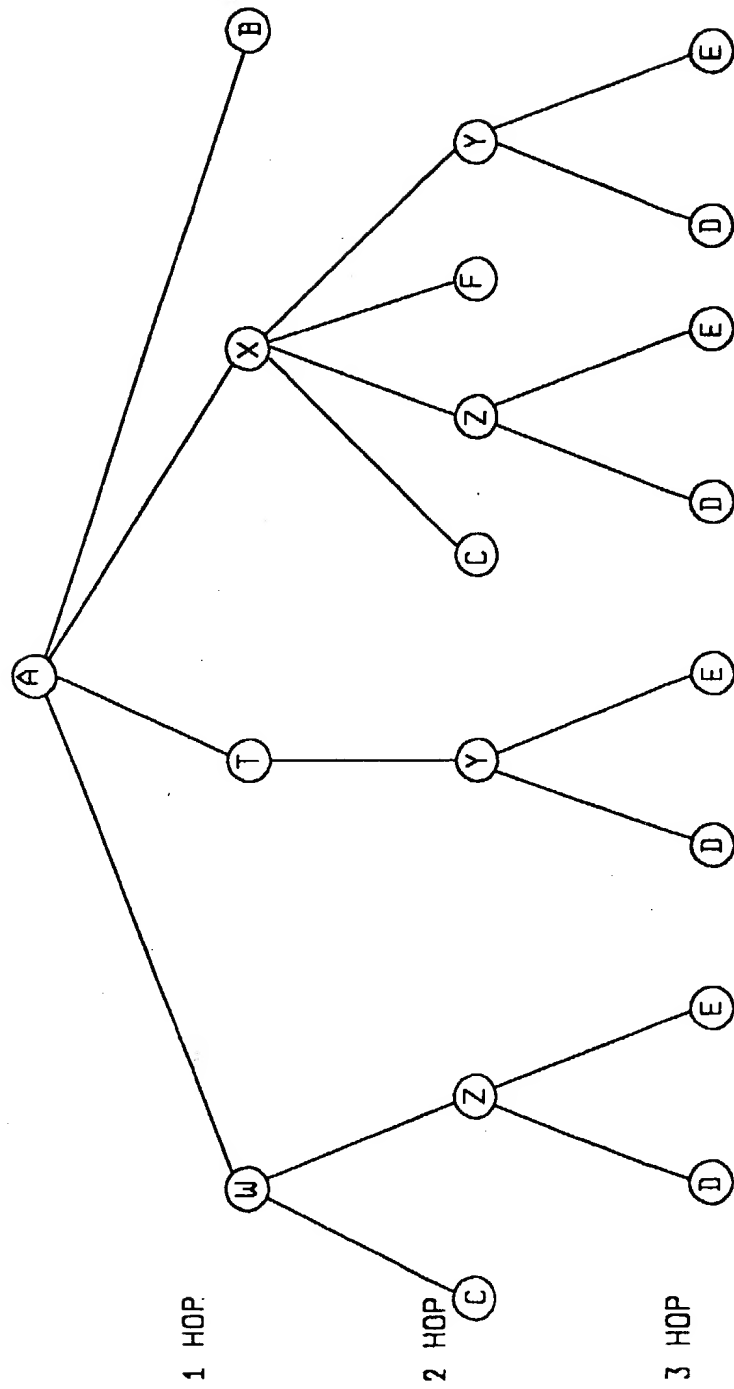


FIGURE 6

SHOULD BE USED TO  
 NODE A TO NODE E - - 00

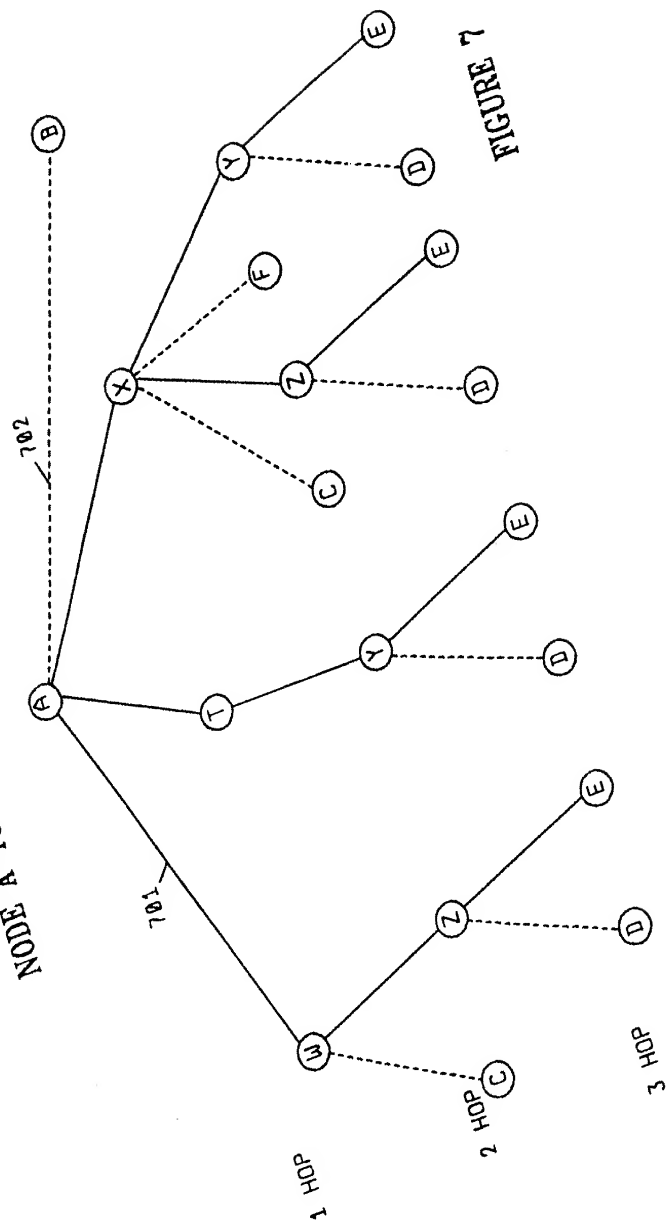


FIGURE 7

## METHOD FOR PRE-SELECTING THE TRANSMISSION LINKS

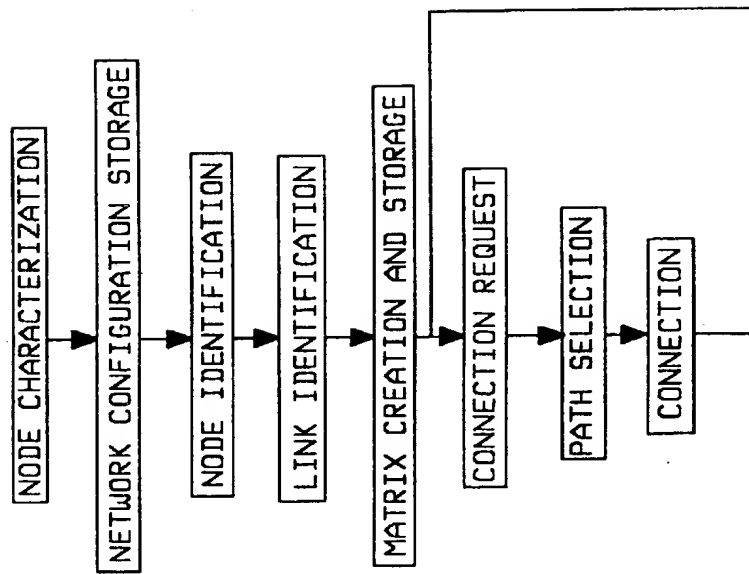


FIGURE 8



European Patent  
Office

## EUROPEAN SEARCH REPORT

Application Number  
EP 93 48 0104

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X	COMPUTER NETWORKS. vol. 1 , 1977 , AMSTERDAM NL pages 155 - 174 L.KLEINROCK ET AL 'HIERARCHICAL ROUTING FOR LARGE NETWORKS' * paragraph 1 * * paragraph 2 *	1,5	H04L12/56
A	IEEE TRANSACTIONS ON COMPUTERS vol. 38, no. 8 , August 1989 , NEW YORK US pages 1059 - 1074 XP47576 W.T.TSAI ET AL 'AN ADAPTIVE HIERARCHICAL ROUTING PROTOCOL' * paragraph I * * paragraph II.A *	1-8	
A	IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION. vol. 7, no. 8 , October 1989 , NEW YORK US pages 1243 - 1252 XP126342 V.R.SAKSENA 'TOPOLOGICAL ANALYSIS OF PACKET NETWORKS' * paragraph I * * paragraph II * * paragraph III *	1-8	
			TECHNICAL FIELDS SEARCHED (Int.Cl.6)
			H04L
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 2 February 1994	Examiner Canosa Areste, C
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document		T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons * : member of the same patent family, corresponding document	